

Learning with many experts: model selection and sparsity

Rafael Izbicki and Rafael Bassi Stern *

May 15, 2014

This is the pre-peer reviewed version of the following article: Izbicki, R., Stern, R. B. “ Learning with many experts: Model selection and sparsity.” Statistical Analysis and Data Mining 6.6 (2013): 565-577., which has been published in final form at <http://onlinelibrary.wiley.com/doi/10.1002/sam.11206/full>

Abstract

Experts classifying data are often imprecise. Recently, several models have been proposed to train classifiers using the noisy labels generated by these experts. How to choose between these models? In such situations, the true labels are unavailable. Thus, one cannot perform model selection using the standard versions of methods such as empirical risk minimization and cross validation. In order to allow model selection, we present a surrogate loss and provide theoretical guarantees that assure its consistency. Next, we discuss how this loss can be used to tune a penalization which introduces sparsity in the parameters of a traditional class of models. Sparsity provides more parsimonious models and can avoid overfitting. Nevertheless, it has seldom been discussed in the context of noisy labels due to the difficulty in model selection and, therefore, in choosing tuning parameters. We apply these techniques to several sets of simulated and real data.

1 Introduction

In many situations, getting reliable labels in a dataset is very expensive and therefore assigning highly trained experts to do such tasks is undesirable. In other situations, even trained experts disagree about the labels of the data. Cases like these include spam detection, diagnosis of patients based on images and morphological classification of galaxies [Lintott et al. \[2008\]](#). Although it might be expensive to train experts to a degree one can trust their labels [Richards et al. \[2012\]](#), systems such as Amazon Mechanical Turk [Schulze et al. \[2011\]](#) allow each sample unit to be classified by many (not necessarily perfect) experts by a reasonably small cost. These experts do not have to be people. For instance, they can be different cheap screening tests in a medical problem [Johnson et al.](#) In these situations, it is desirable to have methods that can detect how reliable each expert is and use this information not only to detect the adequate labels of the data but also to train accurate classifiers to predict new data [Attenberg et al. \[2012\]](#). These methodologies are usually called crowdsourcing methods.

*Department of Statistics, Carnegie Mellon University

Here we focus on predicting binary variables, even though similar ideas can be used in the case of predicting a categorical variable with more than two labels.

The most common approach to deal with multiple experts is to first consider a majority vote scheme to input the labels of each sample unit. Such procedure is known to be suboptimal in many situations [Yan et al., 2010b]. Many other approaches have been proposed recently. While some of them are based on a two step procedure of first trying to find the true labels in the data and then training classifiers based on them Chittaranjan et al. [2011], Karger et al. [2011] others do these tasks simultaneously, that is, the classifier is trained by assuming that the labels from the experts may be incorrect Raykar et al. [2010], Yan et al. [2010b]. We follow the latter approach, even though the model selection technique we propose works for the former method as well.

Also, many of the existing methods are essentially algorithm-based Donmez et al. [2009]. However, a significant amount of the recent methods consist of probabilistic approaches to this problem, in which the unobserved true label is modeled as a latent variable Welinder and Perona [2010], Ipeirotis et al. [2010], Raykar et al. [2010], Yan et al. [2010b,a], Kajino et al. [2012]. In the latter case, the parameters of the model are usually estimated through the Expectation Maximization Algorithm McLachlan and Krishnan [2008]. This approach has roots on [Dawid and Skene, 1979]. However, less emphasis has been given to develop ways of comparing these different models. Since the usual techniques for model selection depend on observing the real labels of the data, they cannot be used in this case. Lam and Stork [2003] discusses how to find good models when only one annotator is available. Here we extend some of these results and relax some of the assumptions made. We take a predictive approach: by good models we mean models that have low predictive errors.

The literature also lacks on methods that can build sparse (in terms of coefficients of the model related to the features) classifiers in crowdsourcing methods. Sparsity is a useful tool when trying to build classifiers that have good generalization properties, that is, that do not suffer from overfitting. Moreover, many common models used for crowdsourcing have a number of parameters that grow with both the number of experts and samples. Having too many parameters can increase the prediction error substantially. Introducing sparsity on such classifiers leads to more parsimonious models that potentially have better performance. Bayesian methods such as the one used by Raykar et al. [2010] can lead to shrinkage of the coefficients and therefore to better prediction errors, however it is not clear how to choose prior hyperparameters on them when one aims at good prediction errors. Sparse methods are also valuable because they can reduce costs: for example, in new samples a smaller number of variables have to be measured.

In Section 2 we develop a method for model selection. In Section 3 we present a model which allows sparse solutions. We also show how to fit the model parameters for a fixed value of the parameter which specifies the amount of sparsity. Section 4 provides applications of both the model selection technique and the sparse model we propose. In particular, we use our model selection technique to select the tuning parameter which induces the classifier with best predictive errors

2 Model Selection

Assume there are d experts that label n sample units. For each of these units, we measure k features. X_{ij} is the j -th feature of the i -th unit. X_i denotes the vector of all features for the i -th sample. $Y_{ir} \in \{0, 1\}$ denotes the label attributed by the r -th expert (annotator) to the i -th sample. $Z_i \in \{0, 1\}$ is the unobserved variable which corresponds to the appropriate label for the i -th sample unit. Table 1 contains a summary of this notation.

Table 1: Model's random variables

True Labels	Experts' Labels			Features		
Z_1	Y_{11}	\dots	Y_{1d}	X_{11}	\dots	X_{1k}
\vdots	\vdots	\ddots	\vdots	\vdots	\ddots	\vdots
Z_n	Y_{n1}	\dots	Y_{nd}	X_{n1}	\dots	X_{nk}

We assume one wishes to find a classifier which minimizes the 0-1 loss. That is, one is interested in finding a classifier that has small probability of making a mistake on a new sample. In this case, many techniques of model selection rely on calculating empirical errors on a test data set [Hastie et al. \[2001\]](#). When using noisy labels, the empirical error is unavailable and this strategy cannot be directly applied. In order to overcome this difficulty, we introduce a score which is closely related to the empirical error. The reliability of this score does not depend on assuming that the data is generated according to the model in Section 3.

Our score is based on splitting the data into a training set and testing set, $(X_i^{test}, Y_i^{test})_{1 \leq i \leq n'}$. If this cannot be done due to a small sample size, one can use a cross validated version of it [Hastie et al. \[2001\]](#). Consider a set of models, Λ . For example, this set can be composed of all models generated by different λ values in the model presented in Section 3. It could also be the set of models fit with different subsets of the features or it could even contain different models such as those obtained using majority vote to input the labels or models such as in [Raykar et al. \[2010\]](#).

For each $\lambda \in \Lambda$, we train the model using the training set. Call z^λ the classifier λ obtained from the training data. Through model selection we wish to find $\lambda \in \Lambda$ with the smallest risk. Define the risk of λ as $R(\lambda) = E[I(Z \neq z^\lambda(X))]$, that is, the probability of a new sample unit being misclassified by the classifier λ . Let n be the sample size in the testing data set. We use z_i^λ as shorthand for $z^\lambda(X_i^{test})$. The (in practice incalculable) empirical risk of model λ is $\hat{R}(\lambda) = \frac{1}{n'} \sum_{i=1}^{n'} \mathbb{I}(z_i^\lambda \neq Z_i)$. For each $\lambda \in \Lambda$, we score how bad z^λ performs through \hat{S} ,

$$\hat{S}(\lambda) = \frac{1}{n'} \sum_{i=1}^{n'} \frac{1}{d} \sum_{j=1}^d I(z_i^\lambda \neq Y_{i,j}^{test}),$$

and select the model λ^* such that

$$\lambda^* = \arg \min_{\lambda \in \Lambda} \hat{S}(\lambda)$$

We prove λ^* is consistent (in the sense of asymptotically giving the same results as when minimizing the real risk $R(\lambda)$) and provide an upper bound on its rate of convergence. In the following theorems, $VC(\Lambda)$ is the VC-dimension of Λ and D a universal constant defined in [Vaart and Wellner \[2000\]](#).

Assumption 1. For every $i \neq j$, $(X_i, Z_i, (Y_{i,k})_{k=1}^d)$ is independent of $(X_j, Z_j, (Y_{j,k})_{k=1}^d)$.

Assumption 2. Let $\epsilon_j = P(Z \neq Y_j)$ be the imprecision of expert j . $\bar{\epsilon} \equiv \frac{\sum_{j=1}^d \epsilon_j}{d} < \frac{1}{2}$.

Assumption 2 means that the label provided by an expert picked uniformly is better than the flip of a coin. We now consider two additional assumptions and then prove that using \hat{S} to perform model selection works under either of them.

Assumption 3. For all i , $\text{Cov}\left(\frac{\sum_{j=1}^d (1 - 2I(Z_i \neq Y_{i,j}^{test}))}{d}, I(z_i^\lambda \neq Z_i)\right) = 0$.

Assumption 3 holds e.g. when the classifier and expert errors are unrelated, a condition that appears on [Lam and Stork \[2003\]](#).

Assumption 4. For all i and $j \neq j^*$, $\text{Cov}(I(Z_i \neq Y_{i,j}^{test}), I(Z_i \neq Y_{i,j^*}^{test})) = 0$.

Assumption 4 holds e.g. when the errors of every two experts are unrelated. We prove the following Theorems:

Theorem 1. Under Assumptions 1 and 3, if Λ is a VC-Class,

$$P(\sup_{\lambda \in \Lambda} |\hat{S}(\lambda) - (1 - 2\bar{\epsilon})R(\lambda) - \bar{\epsilon}| > \delta) \leq \left(\frac{D\sqrt{n'}\delta}{\sqrt{2VC(\Lambda)}} \right)^{2VC(\Lambda)} e^{-2n'\delta^2}.$$

Theorem 2. Under Assumptions 1 and 4, if Λ is a VC-Class,

$$P(\sup_{\lambda \in \Lambda} |\hat{S}(\lambda) - (1 - 2\bar{\epsilon})R(\lambda) - \bar{\epsilon}| > \frac{1}{4\sqrt{d}} + \delta) \leq \left(\frac{D\sqrt{n'}\delta}{\sqrt{2VC(\Lambda)}} \right)^{2VC(\Lambda)} e^{-2n'\delta^2}.$$

The proofs of these facts are sketched in Appendix B. Thus, Theorem 1 states that, under assumptions 1 and 3, as n increases, with high probability, $\hat{S}(\lambda)$ will not deviate more than roughly $\frac{1}{\sqrt{n'}}$ from $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$ (uniformly). Moreover, under assumption 2, $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$ increases on $R(\lambda)$. Hence, the minimizer of $\hat{S}(\lambda)$ will be close to the minimizer of $R(\lambda)$. Using Theorem 2, the same type of reasoning applies under 4, with the exception that $\hat{S}(\lambda)$ will not deviate more than roughly $\frac{1}{\sqrt{n'}} + \frac{1}{4\sqrt{d}}$ from $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$. Hence, consistency is obtained only if the number of experts also increases. Next section describes how to introduce sparsity on a particular model from the literature. In Section 4 we discuss how to select the tuning parameter for this model using \hat{S} .

3 Model Description and Sparse Fitting

The model we use is described by the following conditions, where we use the same notation as in Table 1:

- (i). $(Z_i)_{i \leq n}$ are conditionally independent given $(X_i)_{i \leq n}$.
- (ii). $Z_i | X_i = x_i \sim \text{Ber}\left(\frac{\exp\{\beta_0 + \sum_{j=i}^k \beta_j x_{ij}\}}{1 + \exp\{\beta_0 + \sum_{j=i}^k \beta_j x_{ij}\}}\right)$.
- (iii). $(Y_{ij})_{i \leq n, j \leq k}$ are conditionally independent given $(Z_i)_{i \leq n}$ and $(X_i)_{i \leq n}$.
- (iv). $P(Y_{ik} \neq Z_i | Z_i = z_i, X_i = x_i) = \frac{1}{1 + \exp\{\alpha_k + \sum_{j=i}^k \gamma_j x_{ij}\}}$.

This model is similar to the one specified in [Yan et al. \[2010b\]](#) with the exception that the γ_j coefficients do not depend on the expert. The model's parameters can be interpreted. The higher an α_k is, the more precise the k -the expert. Also, the γ_j coefficients explain how each feature influences the difficulty in classifying a sample unit. β 's are the coefficients that measure the influence of the covariates on the real response. One implicit assumption is that the influence of each feature is the same for all experts. The number of parameters in this model is more than twice the number of features. Thus, sparse classifiers might improve the prediction error if n is small [Hastie et al. \[2001\]](#).

We choose this model because it is simple enough and yet sufficiently reasonable to be applied to many practical situations. We do not intend to argue that this is the best model in all situations. However, similar ideas of how to introduce sparsity can be used in other models from the literature.

The joint distribution of (Y, Z) given X corresponds to a mixture of products of independent Bernoulli variables. In fact, denoting

$$\mu_i := \frac{\exp\{\beta_0 + \sum_{j=i}^J \beta_j x_{ij}\}}{1 + \exp\{\beta_0 + \sum_{j=i}^J \beta_j x_{ij}\}},$$

the complete likelihood (conditional on the features) is given by

$$\begin{aligned} L(y, z; \theta, x) = \\ \prod_i P(\forall k, Y_{ik} = y_{ik}, Z_i = z_i | X_i = x_i) = \prod_i P(\forall k, Y_{ik} = y_{ik} | Z_i = z_i, X_i = x_i) P(Z_i = z_i | X_i = x_i) = \\ \prod_i \left(\prod_k P(Y_{ik} = y_{ik} | Z_i = z_i, X_i = x_i) \right) P(Z_i = z_i | X_i = x_i) = \prod_i \mu_i^{z_i} (1 - \mu_i)^{1-z_i} \times b_i, \end{aligned}$$

where

$$\begin{aligned} b_i = \prod_k \left[\left(\frac{\exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right)^{y_{ik}} \left(\frac{1}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right)^{1-y_{ik}} \right]^{z_i} \times \\ \times \left[\left(\frac{1}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right)^{y_{ik}} \left(\frac{\exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right)^{1-y_{ik}} \right]^{1-z_i} \end{aligned}$$

(that is, b_i is the joint probability of the experts responses conditional on the true labels and on the explanatory variables) and θ indicates all of the model's parameters. Hence, the (complete) log-likelihood is given by

$$\begin{aligned} l(y, z; \theta, x) = \\ \sum_i \left(\sum_k d_{ik} \log \left(\frac{\exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right) + (1 - d_{ik}) \log \left(\frac{1}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right) \right) + \\ + z_i \log(\mu_i) + (1 - z_i) \log(1 - \mu_i), \end{aligned} \tag{1}$$

where

$$d_{ik} := 1 + 2z_i y_{ik} - z_i - y_{ik}.$$

Traditionally, the (local) maximum of the marginal likelihood (defined as $L(y; \theta, x) = \sum_z L(y, z; \theta, x)$) is found by using the EM algorithm [Dempster et al. \[1977\]](#). We propose to introduce sparsity to this solution. Sparsity reduces the number of parameters we have to estimate, and hence can improve the prediction error. For a comprehensive account of the role of sparsity on prediction problems, the reader is referred to [\[Hastie et al., 2001\]](#). To find a sparse fit of the model, instead of maximizing the marginal likelihood, we introduce a L_1 -penalty in the function. That is, we compute,

$$\arg \sup_{\theta} \left(l(y; \theta) - \lambda \sum_{j=1}^k |\beta_j| - \lambda \sum_{j=1}^k |\gamma_j| \right), \quad (2)$$

for some fixed $\lambda > 0$. $l(y; \theta)$ is the log-likelihood of the observed noisy labels, y . Section 2 indicates how one can pick an optimum value of $\lambda > 0$. Other penalties (e.g., L_2) could also lead to better prediction errors, however L_1 penalty creates sparse solutions (that is, it not only shrinks the coefficients) and, as we will see, is tractable from a computational point of view. In order to solve Equation 2, we will first rephrase it in terms of a Bayesian problem that leads to the same results. Imagine that we assign a prior probability for θ as follows:

$$\pi(\theta) \propto \exp \left(-\lambda \sum_{j=1}^k |\beta_j| - \lambda \sum_{j=1}^k |\gamma_j| \right). \quad (3)$$

The maximum a posteriori estimate (MAP) for θ , given Y and X , corresponds to the solution of Equation 2. Let

$$g(\theta, z) := l(y, z; \theta, x) - \lambda \sum_{j=1}^k |\beta_j| - \lambda \sum_{j=1}^k |\gamma_j|,$$

where $l(y, z; \theta, x)$ is as in Equation 1. To find the MAP estimate we use a MAP-EM algorithm [McLachlan and Krishnan \[2008\]](#). That is, we first initialize θ with some given values. Then, we iterate until convergence:

- (i). **(Expectation step)** Find the expected value of the $g(\theta, Z)$, conditional on the current estimates of the parameters θ and on y_{ij} (denoted by $E[g(\theta, Z)]$).
- (ii). **(Maximization step)** Maximize $E[g(\theta, Z)]$ with respect to θ .

Since $g(\theta, Z)$ is linear in Z , the Expectation step follows directly from calculating

$$E[Z_i | Y_{ik} = y_{ik} \ \forall i, k] = \frac{\mu_i \cdot \exp\{\sum_k y_{ik} * (\alpha_k + \sum_j \gamma_j x_{ij})\}}{\mu_i \cdot \exp\{\sum_k y_{ik} * (\alpha_k + \sum_j \gamma_j x_{ij})\} + (1 - \mu_i) \exp\{\sum_k (1 - y_{ik}) * (\alpha_k + \sum_j \gamma_j x_{ij})\}}.$$

and plugging these values into $g(\theta, Z)$. Denote by $g(\theta, z)$ the expected value of $g(\theta, Z)$. For the Maximization step, observe that

$$\arg \sup_{\theta} g(\theta, z) = \quad (4a)$$

$$\arg \sup_{\gamma' s, \alpha' s} \sum_i \left(\sum_{k \in A_i} d_{ik} \log \left(\frac{\exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right) + (1 - d_{ik}) \log \left(\frac{1}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}} \right) \right) \quad (4b)$$

$$- \lambda \sum_{j=1}^k |\gamma_j| \quad (4c)$$

$$+ \arg \sup_{\beta} \sum_i (z_i \log(\mu_i) + (1 - z_i) \log(1 - \mu_i)) - \lambda \sum_{j=1}^k |\beta_j|. \quad (4d)$$

Hence, we have two independent maximization problems, 4b and 4d. Each of them correspond to solving for Weighted L1-Regularized Logistic Regressions, which is implemented in functions such as `glmnet` [Friedman et al. \[2010\]](#) in R. More details on this are given in Appendix A.

The MAP-EM often converges to different points according to the initialization values. One reason for this is that them MAP-EM is guaranteed to converge only to local maximums. A more important reason is due to a type of non-identifiability [Reilink et al. \[1994\]](#) in the model. The parameters (α, γ, β) and $(-\alpha, -\gamma, -\beta)$ induce the same distribution for the data¹. This is common in mixture models and is known as trivial non-identifiability [ná and Renals \[2000\]](#). Consequently, the likelihood will have two optimizers. In order to choose between these points we assume that, averaging over all experts, the probability of correct classification is larger than 50%. This assumption was discussed in Section 2 and can also be found in [Karger et al. \[2011\]](#). Using this assumption, if the MAP-EM converges to θ , we choose between θ and $-\theta$, selecting the classifier which agrees the most with majority vote.

Next section shows empirical performance of this method and the model selection technique in both simulated and real datasets. In particular we discuss how to use the model selection technique from Section 2 to choose the tuning parameter λ .

4 Experiments

We perform 4 experiments that aim at exploring the two methods proposed (sparsity and model selection). Experiment in 4.1, is completely simulated: we generate the features, real responses and also responses from experts. This allows the Bayes error to be calculated. Experiments 4.2 and 4.3 use data from the UCI repository [Newman et al. \[1998\]](#). These databases only contain features and appropriate labels and, thus, we complement them with simulated responses from hypothetical experts. In 4.1, 4.2 and 4.3 we generate the votes from the experts in three ways:

¹For example, consider there is only one expert and that Z represents if a patient is sick or not. We get the same probability that the expert finds the patient to be sick when the expert has good accuracy and the patient has a high probability of being sick (parameters (α, γ, β)) and when the expert has a bad accuracy and the patient has a small probability of being sick (parameters $(-\alpha, -\gamma, -\beta)$).

- (i). The probabilities of misclassification do not depend on the observed features,
- (ii). The probabilities of misclassification follow the model described in Section 3
- (iii). The probabilities of misclassification do not follow Section 3.

The exact description of how the votes were generated varies and is described in each example.

Experiment 4.4 presents a real data set in which a large set of experts responses (42) is available. Hence, majority vote gives us the real response with high probability. For instance, assuming each expert is correct with probability 70% and the responses from experts are independent, majority vote would get the right label with probability $\approx 99.5\%$. In this example, (i), (ii) and (iii) correspond to taking random subsets of size 3 from the 42 experts and comparing the results we get with the (reliable) majority vote on the 42 experts, as if these were the true labels.

In each experiment, we fit and compare the EM without sparsity (denoted by *EM*), with sparsity (*EM-Sparse*) and a L1-penalized logistic regression on the labels obtained by majority vote (*Majority*). For each of the classifiers obtained, we compute \hat{S} and compare it to \hat{R} (which in practice would not be available), the empirical risk. For the sake of comparison, we also fit a L1-penalized logistic regression on the real labels.

We initialize all the parameters generating Gaussian variables with variance 1. For the α 's and γ 's we pick mean 0. For the β 's, the mean is the corresponding coefficient of the logistic regression fitted through majority vote. In order to avoid local maximums, this procedure was repeated 30 times for each simulation.

4.1 Simulated Data Set

We take sample size 2500. The logit of the probability of each appropriate label being 1 is $\beta_0 + \sum_{j=1}^5 \beta_j x_{ij}$ with $\beta = (-0.1, 1, 0.25, 0.24, -0.3, -0.2)$. $(X_1, X_2, X_3, X_4, X_5)$ follows a multivariate normal with mean $(1, 2, 3, 4, 5)$ and covariance matrix,

$$\begin{pmatrix} 0.50 & 0.10 & 0.25 & 0.10 & 0.10 \\ 0.10 & 0.50 & 0.10 & 0.05 & 0.04 \\ 0.25 & 0.10 & 0.80 & 0.01 & 0.10 \\ 0.10 & 0.05 & 0.01 & 0.40 & 0.10 \\ 0.10 & 0.04 & 0.10 & 0.10 & 0.50 \end{pmatrix}$$

We also include 50 covariates unrelated to the labels generated independently from a standard normal distribution. We generate the experts' responses in the following ways:

- (i) Three experts with misclassification probabilities 0.5, 0.15 and 0.47.
- (ii) Four experts, with misclassification probabilities as in Section 3, with $\alpha = (0, .75, -.1)$ and $\gamma = (.1, .2, -.08, .025, -.065)$.
- (iii) Three experts, with probabilities as in Section 3, with $\alpha = (0, .65, -.12)$ and $\gamma = (.05, .05, -.1, -.1, 0)$ but generating the votes through the square of the covariates.

Figure 1 shows the results of applying the model selection ideas to tune the parameter λ . It also shows the estimated predictive risk (based on the real labels) for *EM*, *EM Sparse* and *Majority*, with an interval with one standard deviation around the mean. The Bayes risk is represented by a horizontal line. It is possible to see that in (i), (ii) and (iii), *EM Sparse* beats the other models. Moreover, plain *EM* does not give satisfactory results. This is because there are many (noninformative) covariates, and hence introducing sparsity becomes crucial. Figures related to (ii) and (iii) also show that \hat{S} is also a useful tool to detect points in which either the EM did not converge: they are the points that have a very different behavior in these curves.

Finally, the results from Table 2 agree with our analysis: using \hat{S} to select among different methods gives the same results as using \hat{R} , that is, when using \hat{S} we also conclude that *EM Sparse* is the best model in this case.

4.2 Ionosphere Data Set

The data set ion holds 351 radar returns which can be “good” or “bad”. There are 34 continuous features. We simulate the expert labels, using at most the 4 first features, in the following ways:

- (i) Five experts with misclassification probabilities 0.6, 0.2, 0.5, 0.4 and 0.4.
- (ii) Four experts, with misclassification probabilities as in Section 3, with $\alpha = (0.7, 0, 1.6, 0.7)$ and $\gamma = (0.3, 0.25, -0.3, 0.1)$.
- (iii) Four experts, with probabilities as in Section 3, $\alpha = (0.7, 0, 1.6, 0.7)$ and $\gamma = (0.3, 0.25, -0.3, 0.1)$, but generating the votes through the square of the covariates.

We use a training set of size 175. Figure 3 shows how we fitted *EM-Sparse* and compares it to the other models. \hat{S} is approximately monotonically increasing with \hat{R} and, thus, the minimizer of \hat{S} has empirical risk close to that of the empirical risk minimizer. In scenario (ii), although λ^* is far from the one which minimizes the empirical risk, their risk is similar. Abrupt variations in the top graphs also indicate cases in which the EM probably did not converge. On the bottom, *EM-sparse* improves on results of both *EM* and *Majority* in all scenarios. Finally, we see from the results of Table 2 that using \hat{S} to select between the different models indicates that *EM Sparse* is the model with smaller estimated predictive risk \hat{R} on these cases.

4.3 Wine Quality Data Set

The data set wine contains 1599 red wines and 11 features such as alcohol content and pH. The wine quality of a sample unit is a number between 0 and 10. We define the appropriate label as 1 if wine quality is greater than 5 and 0, otherwise. We generate the noisy labels, using at most the 5 first features, in the following way:

- (i) Three experts with misclassification probabilities 0.4, 0.3 and 0.5.
- (ii) Four experts, with misclassification probabilities as in Section 3, with $\alpha = (1, -0.5, 2.1, 2.3)$ and $\gamma = (0.25, 0.4, 0.3, 0)$.
- (iii) Three experts, with probabilities as in Section 3, $\alpha = (1, -0.5, 1.2)$ and $\gamma = (0.1, 0.2, -0.2, -0.3, -0.3)$, but generating the votes through the square of the covariates.

We use a training set of size 1000. Figure 2 shows how we fitted *EM-Sparse* and compares it to the other models. Regarding the bottom of the figure, in (i) sparsity reduces the prediction error: both *EM-Sparse* and *Majority* are as good as the model fitted using the real labels and much better than *EM*. In (ii), *Majority* is worse than the other approaches, which have the same performance. In (iii), all models perform close to the one obtained using the real labels. On the top of (iii), λ^* is far from the one which minimizes \hat{R} , but has approximately the same risk. Notice that Table 2 leads us to similar conclusions, hence using model selection ideas introduced here also helps us to decide on what is the best approach, EM or majority vote.

4.4 Astronomy Data Set

The sample units in this data set are galaxies. The label is 1 if the shape of the galaxy is *regular* [Izbicki et al. \[2012\]](#) and 0, otherwise. Each galaxy has been labeled by 42 astronomers from CANDELS team [Kartaltepe et al. \[2011\]](#). For each galaxy, there are 7 features which are summary statistics of the their images. These statistics are further described in [Izbicki et al. \[2012\]](#) and [Lotz et al. \[2004\]](#). The training set is composed of 90 galaxies and the testing set of 85. We perform three experiments, (i), (ii) and (iii), by picking as the noisy labels random subsets of size 3 out of the 42 astronomers. True labels are defined to be the majority vote over the 42 astronomers.

Figure 4 illustrates the procedure of fitting *EM-Sparse* and compares it to *EM* and *Majority*. On the top, minimizing \hat{S} yields the same result as minimizing \hat{R} . On the bottom, *EM-Sparse* and *Majority* have approximately the same performance, close to the performance of the model that was fitted when using the real labels. On the other hand, using *EM* without introducing sparsity leads to slightly worse prediction errors in (iii). We emphasize that the large confidence intervals are due to a small sample size. Hence, it is difficult to get conclusive results of which model is the best in this case. However, the first row of Figure 4 shows in practice that assumptions made in Section 2 for model selection are reasonable for this problem.

Table 2: Values of statistic \hat{S} for the experiments in 4. Bold numbers stand for the minimizer of \hat{S} , * indicates the minimizer of \hat{R} .

Simulated	(i)	0.434*	0.485	0.446
	(ii)	0.459*	0.510	0.483
	(iii)	0.462*	0.463	0.477
Ionosphere	(i)	0.432*	0.477	0.435
	(ii)	0.422*	0.512	0.544
	(iii)	0.323*	0.370	0.398
Wine	(i)	0.432	0.505	0.431*
	(ii)	0.386*	0.389	0.399
	(iii)	0.433	0.450*	0.451
Astronomy	(i)	0.290	0.321	0.286*
	(ii)	0.229*	0.288	0.241
	(iii)	0.323*	0.452	0.335

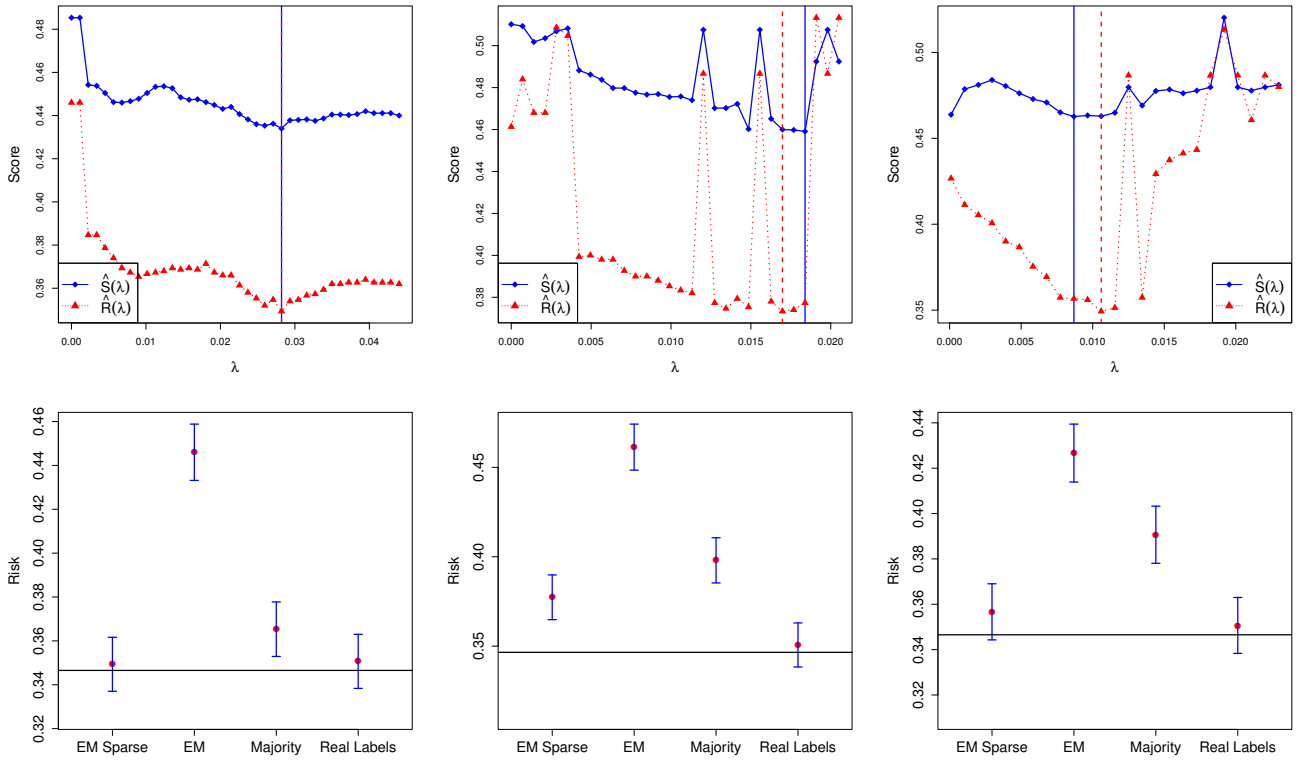


Figure 1: Top: Process of choosing λ using $\hat{S}(\lambda)$ for Simulated Data Set for models (i), (ii) and (iii) respectively. Vertical lines indicate where the minimum is attained for $\hat{S}(\lambda)$ (solid) and for $\hat{R}(\lambda)$ (dashed). Bottom: Estimated prediction errors for each dataset according to each model. Horizontal lines indicate error of the Bayes classifier.

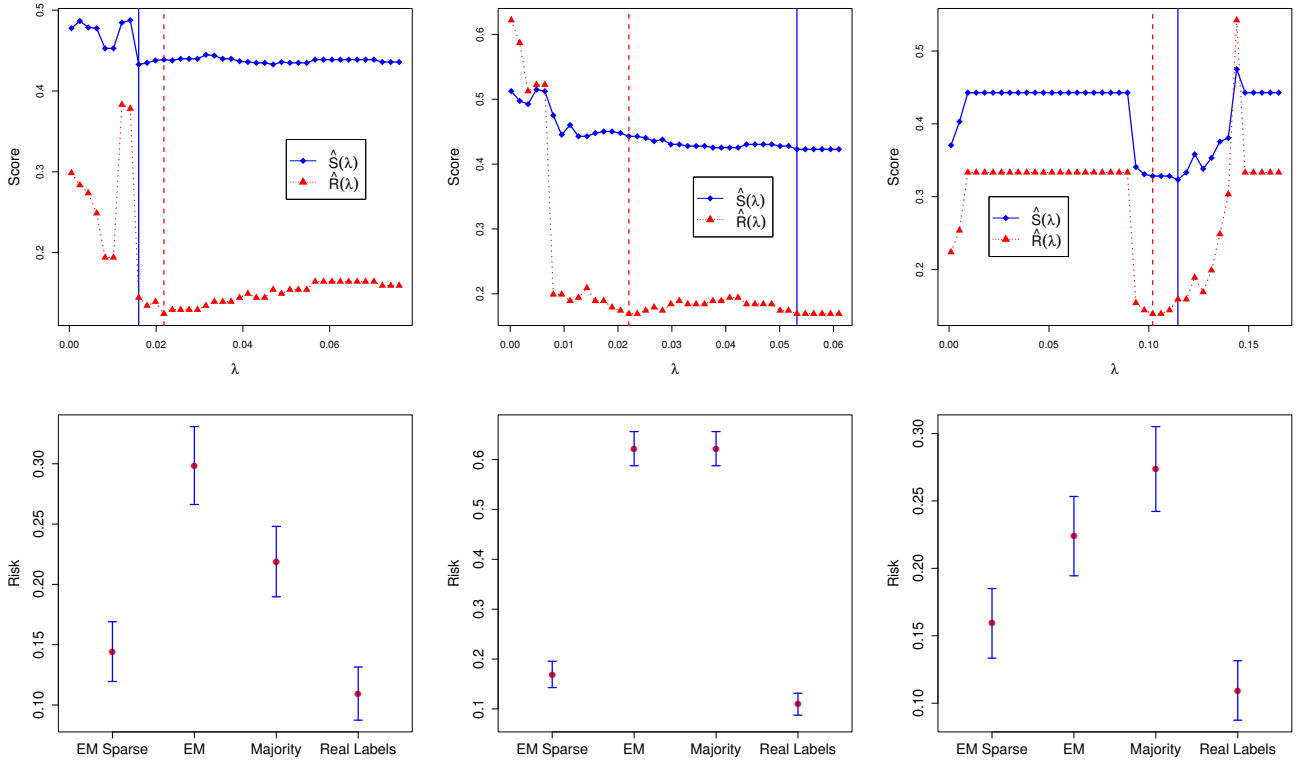


Figure 2: Top: Process of choosing λ using $\hat{S}(\lambda)$ for Ionosphere Data Set for models (i), (ii) and (iii) respectively. Vertical lines indicate where the minimum is attained for $\hat{S}(\lambda)$ (solid) and for $\hat{R}(\lambda)$ (dashed). Bottom: Estimated prediction errors for each dataset according to each model.

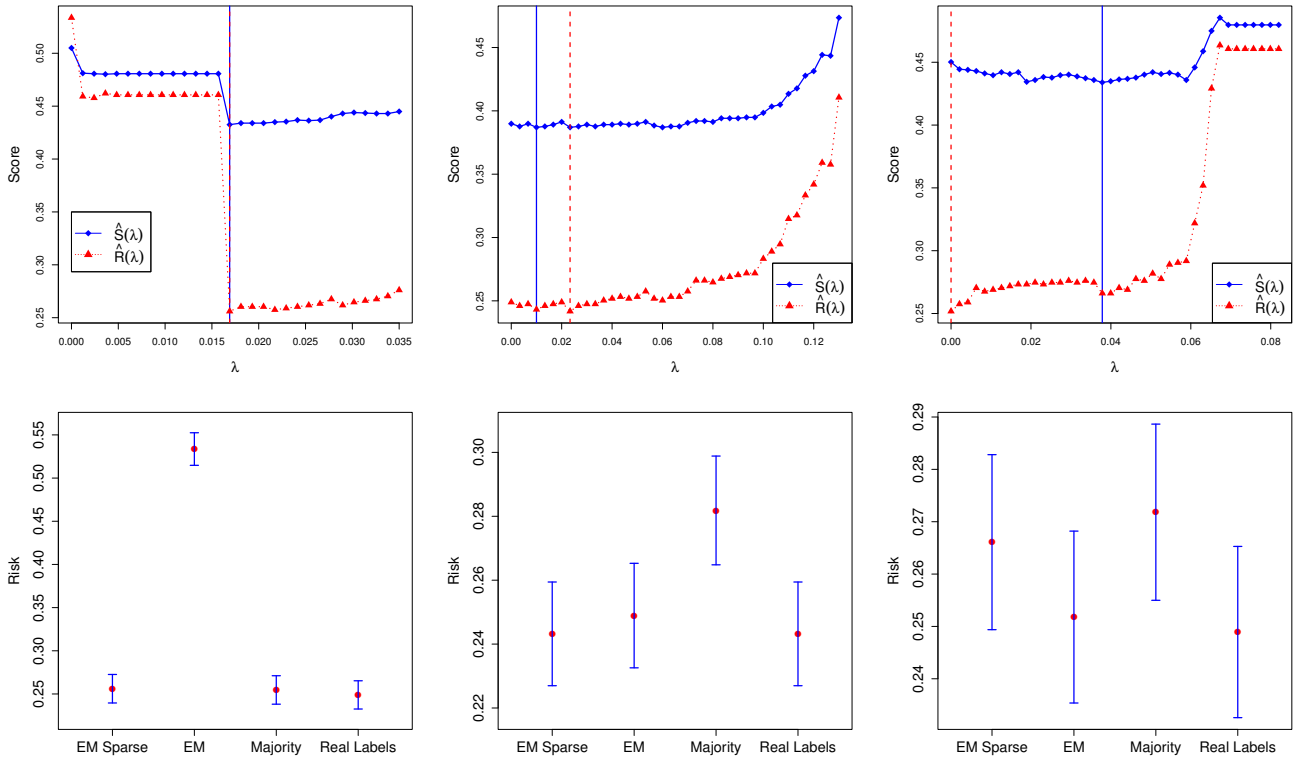


Figure 3: Top: Process of choosing λ using $\hat{S}(\lambda)$ for Wine Data Set for models (i), (ii) and (iii) respectively. Vertical lines indicate where the minimum is attained for $\hat{S}(\lambda)$ (solid) and for $\hat{R}(\lambda)$ (dashed). Bottom: Estimated prediction errors for each dataset according to each model.

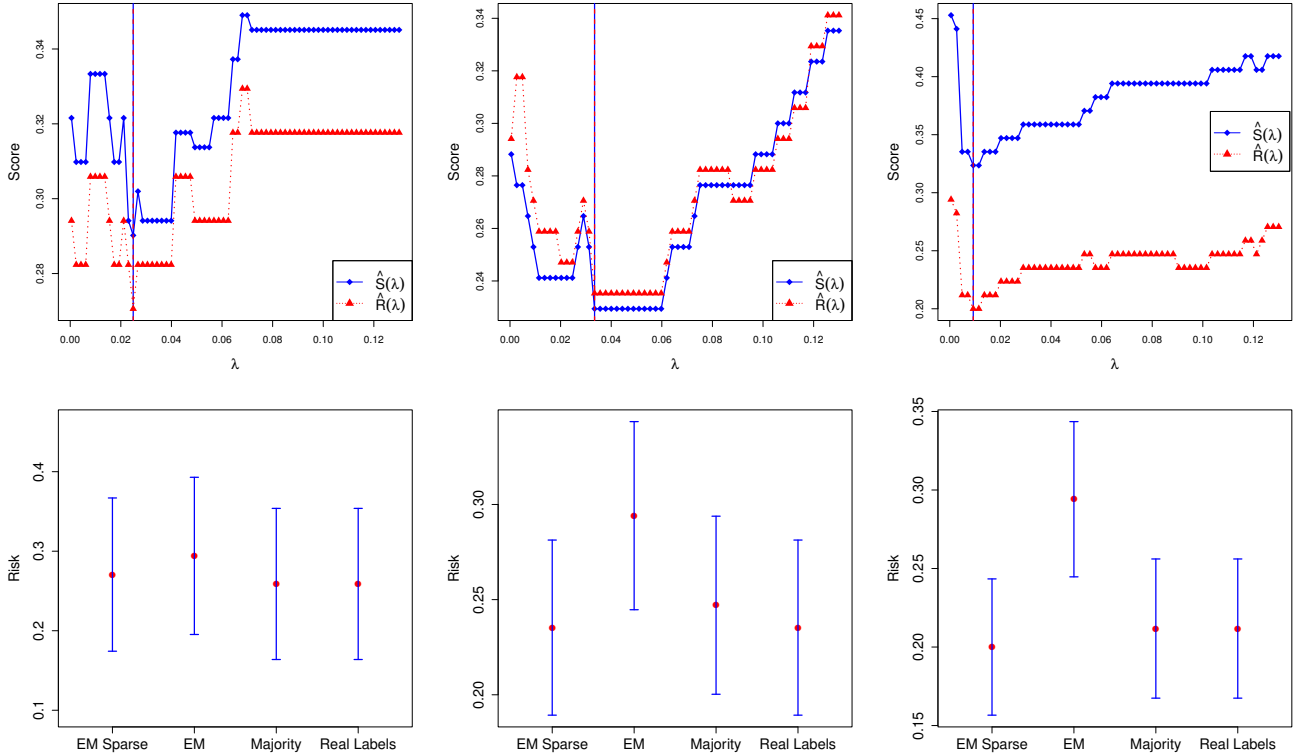


Figure 4: Top: Process of choosing λ using $\hat{S}(\lambda)$ for Astronomy Data Set. Vertical lines indicate where the minimum is attained for $\hat{S}(\lambda)$ (solid) and for $\hat{R}(\lambda)$ (dashed). Bottom: Estimated prediction errors for each dataset according to each model.

5 Conclusions and Future Work

Dealing with noisy labels is a common problem. We show a way one can build classifiers that potentially have better performance than more traditional methods used when true labels are unavailable. The idea behind it is that sparsity is a good way to avoid overparametrizations and therefore creates classification schemes that may have better prediction errors. We also show how model selection can be performed, in particular how one can choose tuning parameters that induce sparsity. The method is based on the introduction of a surrogate function for the estimated risk. Both theoretical and empirical results indicate that the proposed method for model selection works under a fairly large class of problems.

Even though in many situations latent variable models provide big improvements compared to majority vote (see [Yan et al. \[2010b\]](#) and [Raykar et al. \[2010\]](#) for examples of such cases), we saw that in some cases the latter can perform better than the former. Two important reasons of why this happens are 1. The Expectation Maximization Algorithm is sensible to initialization and may converge to local minimums and 2. If the number of experts is large, majority vote can be accurate provided the voters are reasonably good. On the other hand, in such situations latent variable models have too many parameters to be estimated, and hence estimation is more difficult. This is a problem specially if the number of samples is small. However, sparsity can often diminish this problem, leading to estimators that may be better than the ones derived from majority vote procedures. A way to deal with this in practice is to use the proposed model selection technique to compare models built on majority vote labels and on models derived from latent variable models. Performing this procedure in our examples almost always led us to pick the model among *EM Sparse*, *EM* and *Majority* which had the smallest \hat{R} , which is the standard procedure when true labels are available.

Even though we focus on the approach of building models without first estimating the true labels of the data, the ideas of model selection presented are quite general. In fact, even when using the two step procedures (that first find the “true” labels either using majority vote or using fancier methods such as in [Karger et al. \[2011\]](#), and then build classifiers based on the recovered labels), the technique proposed for choosing between models is still valid. An advantage of the latent variables approach over two step procedures is that the first naturally allows partial information from experts to be incorporated when classifying new instances, that is, one can easily calculate $p(z|x, y)$ for new data, even if not all experts observe the data point.

On this paper, we used the same tuning parameter for both γ 's and β 's. As the roles of parameters are of different nature, in practice better performance can be achieved by using two different tuning parameters. This improvement comes at the expense of computational time.

Even though we only introduced sparsity for a specific model, the same arguments can be performed in different situations. For example, one could easily create models in which $P(Y = 1|Z = 1) \neq P(Y = 0|Z = 0)$ by introducing new coefficients. It is also possible to use links different than the logit, and also include dependencies that are not linear in the covariates that were observed. One can also introduce sparsity on approaches from the literature that were already shown to be useful (e.g., [Raykar et al. \[2010\]](#), [Kajino et al. \[2012\]](#)). Our model selection technique helps choosing between these models.

There are also open questions regarding model selection through \hat{S} . Theorem 1 does not hold if any of the assumptions is removed. Necessary conditions for the consistency of minimizing \hat{S} in model selection are unknown. It would also be useful to estimate $R(\lambda^*)$. Theorem 1 shows that $S(\lambda^*)$ is close to $(1 - 2\bar{\epsilon})R(\lambda^*) + \bar{\epsilon}$. Hence, it might be possible to estimate $R(\lambda^*)$ using $S(\lambda^*)$ and an estimator for $\bar{\epsilon}$. Therefore it would be useful to have consistent estimators of $\bar{\epsilon}$. Finally, we use \hat{S} to find a consistent estimator under the 0-1 loss. It remains unknown how to generalize this methodology for other loss functions. For example, in some classification of binary variables, the cost of error depends on the labels.

6 Acknowledgments

The authors are thankful for Peter E. Freeman, Georg M. Goerg, Ann B. Lee, Jennifer M. Lotz, Tiago Mendonça Jeffrey A. Newman, Mauricio Sadinle and Larry Wasserman for the insightful comments. The authors would also like to thank the members of the CANDELS collaboration for providing the proprietary data and annotations for the astronomy data. This work was partially supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico*.

References

- J. Attenberg, P. Melville, F. Provost, and M. Saar-Tsechansky. Selective data acquisition for machine learning. In Balaji Krishnapuram, Shipeng Yu, and Bharat Rao., editors, *Cost-Sensitive Machine Learning*, chapter 5. Chapman & Hall/CRC, Boca Raton, FL, 2012. 1
- G. Chittaranjan, O. Aran, and D. Gatica-Perez. Inferring truth from multiple annotators for social interaction analysis. In *Neural Information Processing Systems (NIPS) Workshop on Modeling Human Communication Dynamics (HCD)*, page 4, 2011. 2
- A.P. Dawid and A.M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28, 1979. 2
- A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. 6
- P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268, 2009. 2
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL <http://www.jstatsoft.org/v33/i01/>. 7, 17

- T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001. [3](#), [5](#), [6](#)
- P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, New York, NY, USA, 2010. ACM. [2](#)
- R. Izbicki, A. B. Lee, P. E. Freeman, and J. A. Newman. Detection of non-regular galaxies at high redshift. Technical report, Department of Statistics, Carnegie Mellon University, 2012. [10](#)
- W. O. Johnson, J. L. Gastwirth, and L. M. Pearson. Screening without a “gold standard”: The hui-walter paradigm revisited. *American Journal of Epidemiology*, 153:921–924. [1](#)
- H Kajino, Y Tsuboi, I Sato, and H Kashima. Learning from crowds and experts, 2012. [2](#), [13](#)
- D.R Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1953–1961. 2011. [2](#), [7](#), [13](#)
- J. S. Kartaltepe, M. Dickinson, D. M. Alexander, E. F. Bell, T. Dahlen, D. Elbaz, S. M. Faber, J. Lotz, D. H. McIntosh, T. Wiklind, B. Altieri, H. Aussel, M. Bethermin, F. Bournaud, V. Charmandaris, C. J. Conselice, A. Cooray, E. Daddi, H. Dannerbauer, R. Davé, J. S. Dunlop, A. Dekel, H. C. Ferguson, N. A. Grogin, H. S. Hwang, R. Ivison, D. Kocevski, A. Koekemoer, D. C. Koo, K. Lai, R. Leiton, R. Lucas, D. Lutz, G. Magdis, B. Magnelli, G. Morrison, M. Mozena, J. Mullaney, J. A. Newman, A. Pope, P. Popesso, A. van der Wel, B. Weiner, and S. Wuyts. GOODS-Herschel & CANDELS: The Morphologies of Ultraluminous Infrared Galaxies at $z \sim 2$. 2011. URL <http://arxiv.org/abs/1110.4057>. [10](#)
- C. P. Lam and D. G. Stork. Evaluating classifiers by means of test data with noisy labels. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003. [2](#), [4](#)
- C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. van D. Berg. Galaxy zoo : Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389, 2008. [1](#)
- J. M. Lotz, J. Primack, and P. Madau. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163, 2004. [10](#)
- G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. 2008. [2](#), [6](#)
- M. Á. Carreira-Perpi n  and S.  . Renals. Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12:141–152, 2000. [7](#)
- C.L. Newman, Blake D.J., and C.J. Merz. UCI repository of machine learning databases, 1998. [7](#)

- V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010. [2](#), [3](#), [13](#)
- M. Reilink, T. Gyllenberg, E. Koski, and M. Verlaan. Non-Uniqueness in Probabilistic Numerical Identification of Bacteria. 31(2):542–548, 1994. [7](#)
- J.W. Richards, D. L. Starr, H. Brink, A.A. Miller, J.S. Bloom, N.R. Butler, J.B. James, J.P. Long, and J. Rice. Active learning to overcome sample selection bias: application to photometric variable star classification. *The Astrophysical Journal*, 2012. [1](#)
- T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader. Exploring task properties in crowdsourcing - an empirical study on mechanical turk. In *ECIS'11*, pages –1–1, 2011. [1](#)
- A. W. Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 2000. [3](#), [19](#)
- P. Welinder and P. Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. *Workshop on Advancing Computer Vision with Humans in the Loop at CVPR*, 2010. [2](#)
- Y. Yan, R. Rosales, G. Fung, and J. Dy. Modeling multiple annotator expertise in the semi-supervised learning scenario. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 674–682, Corvallis, Oregon, 2010a. AUAI Press. [2](#)
- Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, and PA Malvern. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International Conference on Artificial Intelligence and Statistics*, volume 9, pages 932–939, 2010b. [2](#), [5](#), [13](#)

Appendix

A Maximization Step of EM

Here we give more details of how to find the maximum in Equation 4. As we noted, we have two independent maximization problems, 4b and 4d.

The first problem, 4b, can be rewritten as

$$\arg \sup_{\gamma' s, \alpha' s} \sum_{i=1}^{2dn} w(i) \log(\mu'_i),$$

where

$$w_i = \begin{cases} i\text{-th element of the vectorization of the matrix } (d_{jl})_{1 \leq j \leq n, 1 \leq l \leq d} & \text{for } 1 \leq i \leq dn \\ 1 - w_{i-dn} & \text{for } dn + 1 \leq i \leq 2dn \end{cases}$$

and

$$\mu'_i = \begin{cases} i\text{-th element of the vectorization of the matrix } (\mu_{jl})_{1 \leq j \leq n, 1 \leq l \leq d} & \text{for } 1 \leq i \leq dn \\ 1 - \mu'_{i-dn} & \text{for } dn + 1 \leq i \leq 2dn \end{cases}$$

Here,

$$\mu_{ik} = \frac{\exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}{1 + \exp\{\alpha_k + \sum_j \gamma_j x_{ij}\}}$$

This is just a Weighted L1-Regularized Logistic Regression, and can be solved using functions such as glmnet [Friedman et al., 2010] in R. Alternatively, one can directly use algorithms such as Newton-Raphson. Note that if the number of experts is larger than the number of features, using a sparse representation of the matrix can speed up the calculations. The observations related to this maximization problem are

$$\underbrace{1, \dots, 1}_{dn \text{ times}}, \underbrace{0, \dots, 0}_{dn \text{ times}}.$$

The second problem, 4d, can be rewritten as

$$\arg \sup_{\beta} \sum_{i=1}^{2n} w_i \log(\mu'_i) - \lambda \sum_{j=1}^k |\beta_j|,$$

where

$$w_i = \begin{cases} z_i & \text{for } 1 \leq i \leq n \\ 1 - z_{i-n} & \text{for } n + 1 \leq i \leq 2n \end{cases} \quad \mu'_i = \begin{cases} \mu_i & \text{for } 1 \leq i \leq n \\ 1 - \mu_{i-n} & \text{for } n + 1 \leq i \leq 2n \end{cases}$$

This is again a Weighted L1-Regularized Logistic Regression. The observations related to this maximization problem are

$$\overbrace{1, \dots, 1}^{\text{n times}} \overbrace{0, \dots, 0}^{\text{n times}}.$$

B Proofs

Our argument of why \hat{S} is a good measure of performance can be decomposed in three steps. First, we show that the mean of $\hat{S}(\lambda)$ is close to $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$. Next, we prove that, if Λ is a VC-Class, then $\hat{S}(\lambda)$ approaches its mean uniformly on Λ . Finally, since $1 - 2\bar{\epsilon} > 0$ (Assumption 2), the minimizer of $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$ is the same as the minimizer of $R(\lambda)$. From the three steps, we conclude that minimizing $\hat{S}(\lambda)$ approaches minimizing $R(\lambda)$.

We use the following result found to relate $\hat{S}(\lambda)$ to the empirical risk:

Lemma 1. *For all $\lambda \in \Lambda$ it holds that:*

$$\hat{S}(\lambda) = \frac{1}{n'} \sum_{i=1}^{n'} \left(\frac{\sum_{j=1}^d (1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test}))}{d} \right) \mathbb{I}(z_i^\lambda \neq Z_i) + \frac{1}{n'} \sum_{i=1}^{n'} \frac{1}{d} \sum_{j=1}^d \mathbb{I}(Z_i \neq Y_{i,j}^{test})$$

Proof. For any given i and j ,

$$\begin{aligned} \mathbb{I}(z_i^\lambda \neq Y_{i,j}^{test}) &= \mathbb{I}(z_i^\lambda \neq Y_{i,j}^{test}, Z_i = Y_{i,j}^{test}) + \mathbb{I}(z_i^\lambda \neq Y_{i,j}^{test}, Z_i \neq Y_{i,j}^{test}) = \\ &= \mathbb{I}(z_i^\lambda \neq Z_i, Z_i = Y_{i,j}^{test}) + \mathbb{I}(z_i^\lambda = Z_i, Z_i \neq Y_{i,j}^{test}) = \\ \mathbb{I}(z_i^\lambda \neq Z_i)(1 - \mathbb{I}(Z_i \neq Y_{i,j}^{test})) &+ (1 - \mathbb{I}(z_i^\lambda \neq Z_i))\mathbb{I}(Z_i \neq Y_{i,j}^{test}) = \\ &= \mathbb{I}(z_i^\lambda \neq Z_i)(1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test})) + \mathbb{I}(Z_i \neq Y_{i,j}^{test}) \end{aligned}$$

$\hat{S}(\lambda)$ is obtained averaging $\mathbb{I}(z_i^\lambda \neq Y_{i,j}^{test})$ over i and j . The right hand side of the lemma is obtained averaging $\mathbb{I}(z_i^\lambda \neq Z_i)(1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test})) + \mathbb{I}(Z_i \neq Y_{i,j}^{test})$ over i and j . Hence, the proof is complete. \square

Observe that $\frac{1}{n'} \sum_{i=1}^{n'} \frac{1}{d} \sum_{j=1}^d \mathbb{I}(Z_i \neq Y_{i,j}^{test})$ is constant on λ . Hence,

$$\arg \min_{\lambda \in \Lambda} \hat{S}(\lambda) = \arg \min_{\lambda \in \Lambda} \frac{1}{n'} \sum_{i=1}^{n'} \frac{\sum_{j=1}^d (1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test}))}{d} \mathbb{I}(z_i^\lambda \neq Z_i)$$

The model which minimizes $\hat{S}(\lambda)$ minimizes a weighted average of $\mathbb{I}(z_i^\lambda \neq Z_i)$. This is similar to performing model selection through empirical risk minimization, in which the model which minimizes the arithmetic mean of $\mathbb{I}(z_i^\lambda \neq Z_i)$ is chosen.

Lemma 2. *Under assumption 4, for all $\lambda \in \Lambda$ it holds that*

$$\left| E[\hat{S}(\lambda)] - (1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon} \right| \leq \frac{\sigma_{z^\lambda}}{\sqrt{d}},$$

where $\sigma_{z^\lambda}^2 = \text{VAR}[\mathbb{I}(z_i^\lambda \neq Z_i)]$.

Proof. Let $W = \mathbb{I}(z_i^\lambda \neq Z_i)$ and $V_j = 1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test})$. From Cauchy Schwartz inequality it follows that:

$$\left| \text{COV} \left(W, \frac{\sum_j V_j}{d} \right) \right| \leq \sqrt{\text{VAR}[W]} \sqrt{\text{VAR} \left[\frac{\sum_j V_j}{d} \right]} = \sigma_{z^\lambda} \frac{1}{d} \sqrt{\sum_j \text{VAR}[V_j]} \leq \frac{\sigma_{z^\lambda}}{\sqrt{d}}$$

The conclusion follows from noticing that $\left| \text{COV} \left(W, \frac{\sum_j V_j}{d} \right) \right|$ is the left term of the inequality presented. \square

Hence, if we can conclude that $\hat{S}(\lambda)$ is close to its mean, since its mean is close to $(1 - 2\bar{\epsilon})R(\lambda) + \bar{\epsilon}$, we establish that minimizing $\hat{S}(\lambda)$ is close to minimizing $R(\lambda)$. The following result proves that $\hat{S}(\lambda)$ is close to its mean.

Lemma 3. *If Λ is a VC-Class then,*

$$P(\sup_{\lambda \in \Lambda} |\hat{S}(\lambda) - E[\hat{S}(\lambda)]| > \delta) \leq \left(\frac{D\sqrt{n'}\delta}{\sqrt{2VC(\Lambda)}} \right)^{2VC(\Lambda)} e^{-2n'\delta^2}$$

Proof. Using Lemma 1,

$$\hat{S}(\lambda) = \frac{1}{n'} \sum_{i=1}^{n'} \left(\frac{\sum_{j=1}^d (1 - 2\mathbb{I}(Z_i \neq Y_{i,j}^{test}))}{d} \right) \mathbb{I}(z_i^\lambda \neq Z_i) + \frac{1}{n'} \sum_{i=1}^{n'} \frac{1}{d} \sum_{j=1}^d \mathbb{I}(Z_i \neq Y_{i,j}^{test})$$

Define $V_i = \frac{1}{d} \sum_{j=1}^d \mathbb{I}(Z_i \neq Y_{i,j}^{test})$ and $W_i^\lambda = \mathbb{I}(z_i^\lambda \neq Z_i)$. Thus,

$$\hat{S}(\lambda) = \frac{1}{n'} \left(\sum_{i=1}^{n'} W_i^\lambda (1 - 2V_i) + V_i \right)$$

We wish to prove that the central limit theorem holds uniformly on $S[\Lambda] = \{W^\lambda(1 - 2V) + V : \lambda \in \Lambda\}$. Let $N(\mathcal{F}, \epsilon, L^2(Q))$ be the $L^2(Q)$ covering number of a class of functions, \mathcal{F} . Call $R[\Lambda] = \{W^\lambda : \lambda \in \Lambda\}$. Note that, since $|1 - 2V| \leq 1$, for every distribution Q , $N(S[\Lambda], \epsilon, L^2(Q)) \leq N(R[\Lambda], \epsilon, L^2(Q))$. Let $VC(\Lambda)$ be the VC-dimension of Λ . From [Vaart and Wellner \[2000\]](#), $\sup_Q N(R[\Lambda], \epsilon, L^2(Q)) \leq K \cdot VC(\Lambda) (4e)^{VC(\Lambda)} \left(\frac{1}{\epsilon}\right)^{2(VC(\Lambda)-1)}$. Hence, there exists a constant D , such that,

$$P(\sup_{\lambda \in \Lambda} |\hat{S}(\lambda) - E[\hat{S}(\lambda)]| > \delta) \leq \left(\frac{D\sqrt{n'}\delta}{\sqrt{2VC(\Lambda)}} \right)^{2VC(\Lambda)} e^{-2n'\delta^2}$$

\square

Finally, putting together lemmas 2 and 3, we get Theorems 1 and 2.